3D Human Shape Reconstruction from a Polarization Image

Shihao Zou[1], Xinxin Zuo[1], Yiming Qian[2], Sen Wang[1], Chi Xu[3], Minglun Gong[4], and Li Cheng[1]

[1] University of Alberta
[2] Simon Fraser University
[3] School of Automation, China University of Geosciences, Wuhan 430074, China
[4] University of Guelph
{szou2,xzuo,sen9,lcheng5}@ualberta.ca,yimingq@sfu.ca,
xuchi@cug.edu.cn,minglun@uoguelph.ca

Abstract. This paper tackles the problem of estimating 3D body shape of clothed humans from single polarized 2D images, i.e. polarization images. Polarization images are known to be able to capture polarized reflected lights that preserve rich geometric cues of an object, which has motivated its recent applications in reconstructing surface normal of the objects of interest. Inspired by the recent advances in human shape estimation from single color images, in this paper, we attempt at estimating human body shapes by leveraging the geometric cues from single polarization images. A dedicated two-stage deep learning approach, SfP, is proposed: given a polarization image, stage one aims at inferring the fined-detailed body surface normal; stage two gears to reconstruct the 3D body shape of clothing details. Empirical evaluations on a synthetic dataset (SURREAL) as well as a real-world dataset (PHSPD) demonstrate the qualitative and quantitative performance of our approach in estimating human poses and shapes. This indicates polarization camera is a promising alternative to the more conventional color or depth imaging for human shape estimation. Further, normal maps inferred from polarization imaging play a significant role in accurately recovering the body shapes of clothed people.

Keywords: Human Pose and Shape Estimation, Clothed 3D Human body, Shape from Polarization

1 Introduction

Compared to the task of color-image based pose estimation [1–20] that predicts 3D joint positions of an articulated skeleton, human shapes provide much richer information of a human body in 3D and are visually more appealing. It, on the other hand, remains a challenging problem, partly owing to the relative high-dimensional space of human body shapes. The issue is somewhat alleviated by the emerging low-dimensional modelling of human shape, such as SCAPE [21] and SMPL [22], statistical models that are learned from large sets of carefully

scanned 3D body shapes. Based on these low-dimensional human shape representations, a number of end-to-end deep learning methods [23–37] are subsequently developed to estimate human shapes directly from color images. The predicted human shapes, however, are usually naked and lacking in surface details, since e.g. SMPL model is learned from naked human body scans.

Volume-based techniques [38, 39] are widely used in capturing surface details of a clothed human body from a single image. Due to finite computational resource, the estimated human shapes from these methods are usually of low resolution. Saito et al. [40] consider to remedy this by predicting a pixel-aligned implicit surface function that captures more detailed body surface. It however relies on a large training set of detailed 3D human bodies, and the method is still unable to handle complex poses. In the meantime, the methods of [41] and [42] aim to exploit additional geometric cues arising from color image inputs; [41] instead focuses on predicting fine depth maps, and [42] takes on the shading aspect. Unfortunately, accurate and reliable prediction of these geometric cues from a color image is yet another challenging issue - it remains unclear how much one can leverage from such cues. Motivated by these efforts and their limitations, we consider in this paper to work with a new 2D imaging modality, polarization camera, that is known at better preserving fine-scale geometric properties of 3D objects, including human shapes. The intuition comes from basic physics principle: when a light ray reflects off an object, it is polarized and conveys ample geometric cues concerning local surface details of the object, usually represented as surface normal [43, 44]. It may be found to note some biological species are even able to directly perceive light polarization [45, 46], which significantly facilitates their 3D sensing. Empirically, our experiments support that the surface normal maps obtained out of the input 2D polarization images could play an instrumental role in producing accurate and reliable 3D clothed human shapes.

As shown in Fig. 1, our approach, also called SfP, contains two stages. Stage 1 concentrates on predicting accurate surface normal maps from single polarization images[5] by exploiting the associated physics laws as priors. It is then fed into stage 2 in reconstructing the final clothed human shape.

Unlike existing efforts in normal map prediction [41, 42], our approach predicts normal maps by explicitly incorporating the underlying physical laws of polarization imaging, which results in more *reliable* performance. To achieve this, there are two main challenges we need to overcome, namely π-ambiguity of the azimuth angle and the possibly large noise in practical applications. To this end we introduce two ambiguous normal maps $\mathbf{n}_1$ and $\mathbf{n}_2$ (Sec. 3.1) as a physical prior, based on the assumption that the light reflected by human clothing is mostly diffused. Different from [44], each pixel is then classified into one of the three types: the two ambiguous normal maps and background. This is followed by a refinement step to deliver the final surface normal prediction of $\hat{\mathbf{n}}$, that accounts for the possibly-noisy fused normal map output owing to environmental noise and the digital quantization of the polarization camera. Based on the

[5] In this paper, an polarization image has four channels with each channel corresponding to a specific polarizer degree of (0, 45, 90 and 135).

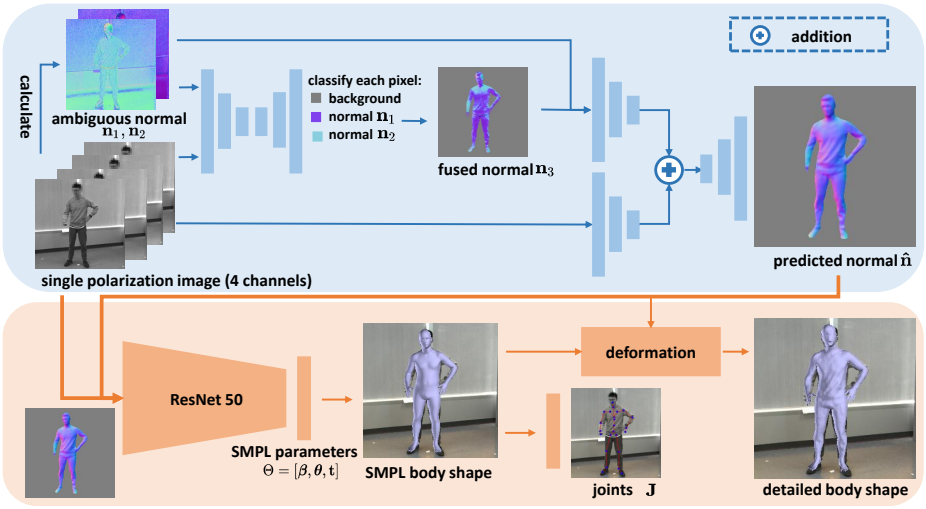

Fig. 1. Given a single polarization image, a two-stage process is executed in our approach. (1) Stage 1, in blue, estimates the surface normal from the polarization image based on the physical assumption that reflected light from an object is polarized. After calculating the two ambiguous normal maps, ($\mathbf{n}_1$, $\mathbf{n}_2$), as physical priors from the polarization image (see Sec. 3.1 for details), image pixels are classified as belonging to either of the two normals or a background, thus obtaining the fused normal $\mathbf{n}_3$. Unfortunately, this normal is often noisy, thud a further step is carried out in regressing a final accurate surface normal $\hat{\mathbf{n}}$, by integrating these physical normal maps and the raw polarization image. (2) Stage 2, in orange, concatenates the polarization image and the surface normal as the input to estimate clothed body shape in two steps. The first step focuses on estimating the parameters of SMPL, a rough & naked shape model parameterized by Θ; the pose (3D joint positions) $\mathbf{J}$ is directly obtained as a by-product of the rigged shape model. The next step deforms the SMPL shape guided by the final surface normal of stage 1, to reconstruct the refined 3D human shape with clothing details.

raw polarization image and output of stage 1, stage 2 concerns the estimation of clothed human shape. It starts from predicting a coarse SMPL shape model, which is then deformed by leveraging the geometric details from surface normal, our stage 1 output, to form the final human shape. Empirically our two-stage pipeline is shown to be capable of accurately reconstructing human shapes, while retaining clothing details such as cloth wrinkles.

To summarize, there are two main contributions in this work. (1) A new problem of inferring high-resolution 3D human shapes from a single polarization image is proposed and investigated. This lead us to curate a dedicated Polarization Human Shape and Pose Dataset (PHSPD). (2) A dedicated deep learning approach, SfP, is proposed[6], where the detail-preserving surface normal maps

[6] Our project website is https://jimmyzou.github.io/publication/2020-polarization-clothed-human-shape

are obtained following the physical laws, and are shown to significantly improve the reconstruction performance of clothed human shapes. Empirical evaluations on a synthetic SURREAL dataset as well as a real-world dataset demonstrate the applicability of our approach. Our work provide sound evidence in engaging 2D polarization camera to estimate 3D human poses and shapes, a viable alternative to conventional 2D color or 3D depth cameras.

2 Related Work

Shape from polarization (SfP) focuses on the inference of shape (normally represented as *surface normal*) from the polarimetric information in the multiple channels of a polarization image, captured under linear polarizers with different angles. The main issue of SfP is angle ambiguity. Previous methods are mainly physics-based that rely on other additional information or assumptions to elucidate the possible ambiguities, such as smooth object surfaces [47], coarse depth map [48, 43] and multi-view geometric constraint [49, 50]. The recent work of [44] proposes to blend physical priors (ambiguous normal maps) with deep learning in uncovering the normal map. Using physical priors as part of the input, deep learning model can then be trained to account for the ambiguity and be noise-resilient. We improve upon [44] by classifying ambiguous normal and background for each pixel, and regressing the normal given the ambiguous and classified physical priors.

3D human pose estimation from single images has been extensively investigated in the past five years, centering around color or depth imaging. Many of the studies [51–57] utilize dictionary-based learning strategies. More recent efforts aim to directly regress 3D pose using deep learning techniques, including CNNs [1–3] and Graph CNNs [58, 59]. In prticulr, several recent efforts [4–12, 12–20] look into a common framework of estimating 2D pose (either 2D joint positions or heatmap), which is then lifted to 3D. Ideas from self-supervised learning [20, 17] and adversarial learning [11, 18] also gain attentions in e.g. predicting 3D pose under additional constraints imposed from re-projection or adversarial losses.

Human shape estimation from single images has drawn growing attentions recently, thanks to development of human shape models of SCAPE and SMPL [21, 22]. These two statistical models learn low-dimensional representations of human shape from large corpus of human body scans. Together with deep learning techniques, it has since been feasible to estimate human body shapes from single color or depth images. Earlier activities focus more on optimizing the SCAPE or SMPL model parameters toward better fitting to various dedicated visual or internal representations, such as foreground silhouette [23–25] and pose [26, 27]. Deep learning based approaches are more commonplace in recent efforts [28–31], which typically learn to predict the SMPL parameters by incorporating the constrains from 2/3D pose, silhouette, as well as adversarial learning losses. [32] takes the body pixel-to-surface correspondence map as proxy representation

and then performs estimation of parameterized human pose and shape. In [33], optimization-based methods [26] and regression-based methods [28] are combined to form a self-improved fitting loop. point cloud is considered as input in [60] to regress SMPL parameters. Instead of single color images, our work is based on single polarization image; rather than inferring coarse human body shape, we aim to recover high-res human shapes.

As for the estimation of clothed human shape, volume-based methods [38–40] are proposed to reconstruct textured body shapes. they unfortunately suffer from the low resolution issue of volumetric representation. Our work is closely related to [42], which combines the robustness of parametric model and the flexibility of free-form 3D deformation in a hierarchical manner. The major difference is, the clothing details of our work are provided by the reliable normal map estimated from the polarization image, whereas the network in [42] deforms depth image by employing the shading information trained on additional data, that are inherently unreliable due to the lack of ground-truth information of surface normal, albedo and environmental lighting. Our work is also related to [41] which recovers detailed human shape from a color image, by iteratively incorporating both rough depth map and estimated surface normal for improved surface details.

3 The Proposed SfP Approach

There are two main stages in our approach: (1) estimate surface normal from a single polarization image; (2) estimate human pose and shape from the estimated surface normal and the raw polarization image, followed by body shape refinement from the estimated surface normal.

3.1 Surface Normal Estimation

The reflected light from a surface mainly includes three components [50], the polarized specular reflection, the polarized diffuse reflection, and the unpolarized diffuse reflection. A polarization camera has an array of linear polarizer mounted right on top of the CMOS imager, similar to the RGB Bayer filters. During the imaging process of a polarization camera, a pixel intensity typically varies sinusoidally with the angle of the polarizer [43]. In this work, we assume that the light reflected off human clothes is dominated by polarized diffuse reflection and unpolarized diffuse reflection. For a specific polarizer angle ϕ_{pol}, the illumination intensity at a pixel with dominant diffuse reflection is

$$\mathrm{I}(\phi_{\text{pol}}) = \frac{\mathrm{I}_{\max} + \mathrm{I}_{\min}}{2} + \frac{\mathrm{I}_{\max} - \mathrm{I}_{\min}}{2} \cos(2(\phi_{\text{pol}} - \varphi)). \tag{1}$$

Here φ is the azimuth angle of surface normal, $\mathrm{I}_{\max}$ and $\mathrm{I}_{\min}$ are the upper and lower bounds of the illumination intensity. $\mathrm{I}_{\max}$ and $\mathrm{I}_{\min}$ are mainly determined by the unpolarized diffuse reflection, and the sinusoidal variation is mainly determined by the polarized diffuse reflection. Note that there is π-ambiguity in the azimuth angle φ in Eq. (1), which means that φ and $\pi + \varphi$ will result in the

same illumination intensity of the pixel. As for the zenith angle θ, it is related to the degree of polarization ρ, where

$$\rho = \frac{\mathrm{I}_{\max} - \mathrm{I}_{\min}}{\mathrm{I}_{\max} + \mathrm{I}_{\min}}. \tag{2}$$

According to [47], when diffuse reflection dominates, the degree of polarization ρ is a function of the zenith angle θ and the refractive index n,

$$\rho = \frac{(n - \frac{1}{n})^2 \sin^2\theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}}. \tag{3}$$

In this paper, we assume the refractive index $n = 1.5$ since the material of human clothes is mainly cotton or nylon. With n known, the solution of θ in Eq. (3) is a close-form expression of n and ρ.

Taking into account the π-ambiguity of φ, we have two possible solutions to the surface normal for each pixel, that form the physical priors. We propose to train a network to classify each pixel into three categories: background, ambiguous normal $\mathbf{n}_1(\varphi, \theta)$ and ambiguous normal $\mathbf{n}_2(\pi + \varphi, \theta)$ with probability p_0, p_1, and p_2 respectively. Then we have the fused normal as follows,

$$\mathbf{n}_3 = (1 - p_0) \cdot \frac{p_1\mathbf{n}_1 + p_2\mathbf{n}_2}{\|p_1\mathbf{n}_1 + p_2\mathbf{n}_2\|_2}, \tag{4}$$

where $(1 - p_0)$ is a soft mask of the foreground human body. Unfortunately, due to the environmental noise and the digital quantization of camera in real-world applications, the fused normal map $\mathbf{n}_3$ is noisy and non-smooth. Thus taking the fused noisy normal as an *improved* physical prior, a denoising network is further trained to take both the polarization image and the physical priors $(\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3)$ as input, and to produce a smoothed normal $\hat{\mathbf{n}}$. The loss function for normal estimation consists of the cross entropy (CE) loss of classification and the L1 loss of the cosine similarity,

$$L_n = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left[\lambda_c \mathrm{CE}(y^{i,j}, p^{i,j}) + \lambda_n (1 - \langle \hat{\mathbf{n}}^{i,j}, \mathbf{n}^{i,j} \rangle) \right], \tag{5}$$

where λ_c and λ_n are the weights of each loss, $y^{i,j}$ is the label indicating which category the pixel (i, j) belongs to, and $\langle \hat{\mathbf{n}}^{i,j}, \mathbf{n}^{i,j} \rangle$ denotes the cosine similarity between the predicted and target normal vectors of pixel (i, j). Note that the category label $y^{i,j}$ is created by discriminating whether the pixel is background or which ambiguous normal has higher cosine similarity with the target normal. λ_c and λ_n is 2 and 1 respectively in our experiment.

3.2 Human Pose and Shape Estimation

To start with, the SMPL [22] representation is used for describing 3D human shapes, which is a differentiable function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^{6,890 \times 3}$ that outputs a

triangular mesh with 6,890 vertices given 82 parameters $[\boldsymbol{\beta}, \boldsymbol{\theta}]$. The shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is the linear coefficients of a PCA shape space that mainly determines individual body features such height, weight and body proportions. The PCA shape space is learned from a large dataset of body scans [22]. $\boldsymbol{\theta} \in \mathbb{R}^{72}$ is the pose parameter that mainly describes the articulated pose, consisting of one global rotation of the body and the relative rotations of 23 joints in axis-angle representation. Finally, our clothed body shape is produced by first applying shape-dependent and pose-dependent deformations to the template pose, then using forward-kinematics to articulate the body shape back to its current pose, and deforming the surface mesh by linear blend skinning. $\mathbf{J} \in \mathbb{R}^{24\times 3}$ are the 3D joint positions that can be obtained by linear regression from the output mesh vertices.

In addition to the SMPL parameters, we also need to predict the global translation $\mathbf{t} \in \mathbb{R}^3$. Thus for the task of human pose and shape estimation, the output vector is of 85-dimension, $\hat{\Theta} = [\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{t}}]$. Given $\hat{\Theta}$, we can also obtain the predicted 3D joint positions $\hat{\mathbf{J}}$. To this end, the loss function is defined as

$$L_s = \lambda_\beta \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2 + \lambda_\theta \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 + \lambda_t \|\mathbf{t} - \hat{\mathbf{t}}\|_2^2 + \lambda_J \|\mathbf{J} - \hat{\mathbf{J}}\|_2^2, \tag{6}$$

where λ_β, λ_θ, λ_t and λ_J are weights of each component in the loss function, which are fixed to 0.2, 0.5, 100, and 3, respectively.

The reconstructed SMPL human shape thus far is naked 3D shape and lacking fine surface details. Our goal is to refine this intermediate naked shape under the guidance of our smoothed surface normal estimate. It is carried out as follows. The SMPL body shape is rendered on the image plane to form a base depth map. The technique in [61] is then engaged here to obtain an optimized depth map from the predicted surface normal and the base depth map. It is carried out under three constraints: first, the predicted normal should be perpendicular to the local tangent of the optimized depth surface; second, the optimized depth should be close to the base depth; Third, a smoothness constraint is enforced on nearby pixels of the optimized depth map. This depth map is obtained as a solution of a linear least-squares system. Weights of the normal term, the depth data term, and the smoothness term are empirically set to 1.0, 0.06, and 0.55, respectively. Finally, our clothed body shape is produced by upsampling & deforming the SMPL mesh according to the Laplacian of the optimized depth map.

3.3 Polarization Human Pose and Shape Dataset

To facilitate empirical evaluation of our approach in real-world scenarios, a home-grown dataset is curated, referred as Polarization Human Shape and Pose Dataset, or PHSPD. A complete description of this PHSPD dataset is provided in [62]. In data Requisition stage, a system of four soft-synchronized cameras are engaged, consisting of a polarization camera and three Kinects V2, with each Kinect v2 having a depth and a color cameras. 12 subjects are recruited in data collection, where 9 are male and 3 are female. Each subject performs 3 different

groups of actions (out of 18 different action types) 4 times, plus an addition period of free-form motion at the end of the session. Thus for each subject, there are 13 short videos (of around 1,800 frames per video with 10-15 FPS); the total number of frames for each subject amounts to 22K. Overall, our dataset consists of 287K frames, each frame here contains a synchronized set of images - one polarization image, three color and three depth images.

The SMPL shape parameters and the 3D joint positions of a body shape are obtained from the image collection of current frame as follows. For each frame, its initial 3D pose estimation is obtained by integrating the Kinect readouts as well as the corresponding 2D joint estimation from OpenPose [63] across the depth and color sensors. Then the body shape, i.e. parameters of the SMPL model, is estimated as optimal fit to the initial pose estimate [26]. The 3D point cloud of body surface acquired from three depth cameras are now utilized in our final step, resulting in the estimation of refined body shape with clothing details [64], by iteratively minimizing the distance of SMPL shape vertex to its nearest point of the 3D point cloud. Exemplar clothed human shapes are shown in Fig. 2.

Fig. 2. Exemplar 3D poses and SMPL shapes in the real-world PHSPD dataset. We render the SMPL shape on four images (one polarization image and three-view color images) and we also show the pose in 3D space.

4 Empirical Evaluations

Empirical evaluations are carried out on two major aspects. (1) For normal estimation, we report the mean angle error (MAE), which measures the angle between the target and estimated normal map, $e_{\text{angle}} = \arccos(\langle \mathbf{n}^{i,j}, \hat{\mathbf{n}}^{i,j} \rangle)$ for

pixel (i, j), where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. (2) For human pose and shape estimation, we report the mean per joint position error (MPJPE) and the 3D surface distance error. MPJPE is defined as the average distance between predicted and annotated joints of the test samples. In both SURREAL and PHSPD datasets, there are 24 annotated 3D joints. We also report the MPJPE for 20 joints by removing the hand and foot joints. The 3D surface error measures the distance between the predicted mesh and the ground truth mesh, by averaged distance of vertex pairs, as follows: for each vertex of the human body mesh, its closest vertex in ground truth mesh is identified to form its vertex pair; the average distance of all such vertex pairs is then computed.

For the real-world PHSPD dataset, subject 4 is chosen to form the validation set (23,786 samples); the test set contains those of subjects 7, 11, and 12 (69,283 samples); the train set has everything else (186,746 samples).

We also demonstrate the effectiveness of our SfP approach on SURREAL [29], a synthetic dataset of color images rendered from motion-captured human body shapes. Polarization images can be synthesized using color and depth images (details are in supplementary material). We choose subset "run1" and select one frame with a gap of ten frames. Finally, the train set has 245,759 samples, validation set has 14,528 samples and test set has 52,628 samples.

4.1 Evaluation of Surface Normal Estimation

In this task, our approach is compared with a recent work *Physics* [44], a traditional method *Linear* [65], and three ablation variants of our method as baselines. *Ours (color image)* uses only color image for estimating the normal map. *Ours (no physical priors)* does not incorporate the ambiguous normal maps as the physical priors and employs the polarization image as the only input. *Ours (no fused normal)* is similar to Physics [44], in which we use the two ambiguous normal maps as the only priors, discarding the fused normal maps.

	SURREAL	PHSPD
Linear [65]	20.03	34.97
Physics [44]	7.45	21.45
ours (color image)	19.49	25.02
ours (no physical priors)	13.89	24.71
ours (no fused normal)	7.43	21.65
ours	**7.10**	**20.75**

Table 1. Comparison of surface normal estimation evaluated in MAE. The competing methods include *Linear* [65], *Physics* [44], *ours*, and three ablation variants of our method.

Through both the quantitative results of Table 1 and the visual results of Fig. 3, it is observed that our method has consistently outperforms the state-

of-the-art surface normal prediction methods [44, 65] in both SURREAL and PHSPD datasets. The poor performance of [65] may be attributed to its unrealistic assumption of noise-free environment in the captured images. Let us look at the three ablation baselines of our approach: using only color images delivers relatively similar performance to that of removing physical priors when compared in PHSPD. Intuitively, it is challenging for neural networks to encode information of ambiguous normal maps (physical priors) directly from raw polarization images. Therefore, removing the physical priors results in similar performance to that of using only color images. [44] and *ours (no fused normal)* both utilize ambiguous normal as a physical prior, thus produce similar results. By incorporating the fused normal which discriminates the ambiguity of azimuth angle estimation, the results of our full-fledged approach surpasses those of [44].

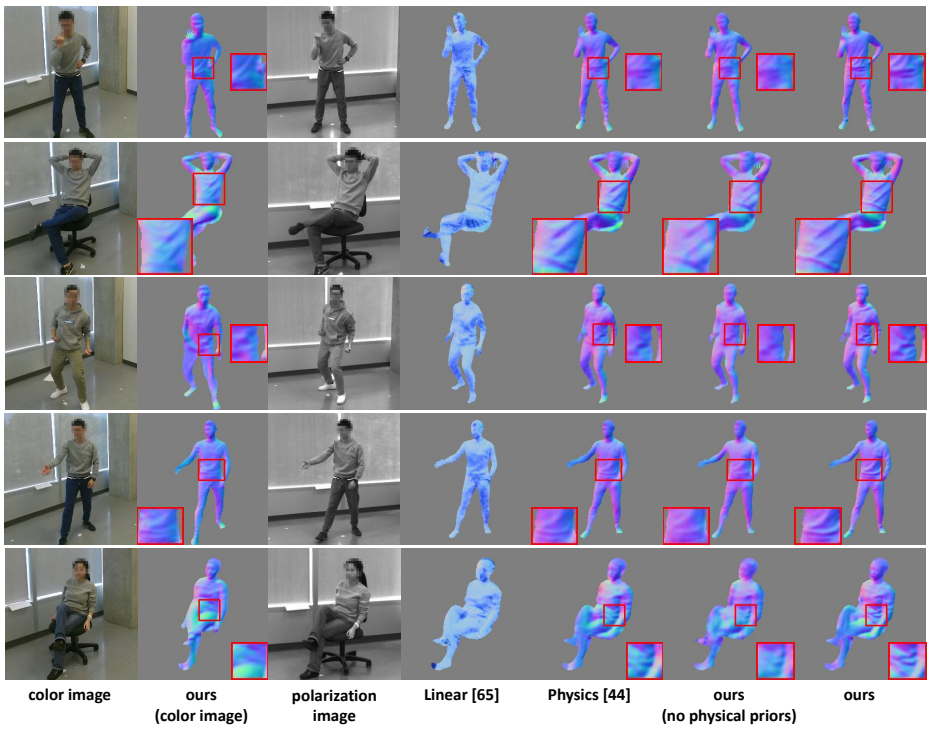

Fig. 3. Exemplar results of normal map prediction by five competing methods: [65, 44], *ours (no physical priors)*, *ours (color image)*, and *ours*. Original color images and polarization images are shown in the first and third column with pixelated faces.

4.2 Evaluation of Pose and Shape Estimation

The focus of this section is qualitative and quantitative evaluations on estimating poses & SMPL shapes, as well as our final estimation of clothed human shapes.

In pose estimation, it is of interest to inspect the effect of engaging surface normal maps in our SfP approach. Besides our SfP approach, the competing methods consist of *HMR* [28] and a ablation variant of SfP, *ours (w/o normal)*. The latter is obtained by engaging only the polarization image, without considering normal map estimation. Since *HMR* is trained on single color images, it is re-trained using the first three channels of a polarization image. In addition to *HMR* that works on color images, for fair comparison, *HMR* is also re-trained on the polarization images of our PHSPD dataset, as *HMR (polarization)*. From Table 2, it is observed that our method produces the lowest MPJPE values of all competing methods; the results of *ours (w/o normal)* is comparable to those of *HMR (polarization)*. The quantitative results confirm that the polarization images is capable of producing accurate estimation of human poses. Moreover, the visual results in Fig. 4 provide qualitative evidence that further performance gain is to be expected, when we have access to the normal maps. Similar observation is again obtained in Table 3, when quantitative examination is systematically carried out over w/ and w/o estimated normal map, on color and polarization images, in both datasets. Note the performance gain is particularly significant for polarization images, which may attribute to the rich geometric information encoded in the normal map representation. On color images, there is still noticeable improvement, also less significant. Our explanation is that the normal maps estimated from color images are not as reliable as those obtained from the polarization image counterparts.

	SURREAL	PHSPD	
	GT-t	GT-t	Pred-t
HMR [28]	116.68/136.32	82.96/91.46	-
HMR (polarization)	-	77.57/88.74	97.24/106.20
ours (w/o normal)	83.43/94.00	84.44/96.42	93.38/104.48
ours	**67.25/75.94**	**66.32/74.46**	**74.58/81.85**

Table 2. Quantitative evaluations using MPJPE evaluation metric on both SURREAL and PHSPD datasets. The unit of the error is millimeter. GT-t means the camera translation is known and Pred-t means the predicted camera translation is used to compute the joint error. We report the MPJPE results of 20/24 joints, which removes two hand and two foot joints following similar settings of previous work [28, 66].

To evaluate the effectiveness of our approach on clothed human shape recovery, the state-of-the-art methods on human surface reconstruction from single color images are recruited. They are *PIFu* [40], *Depth Human* [41] and *HMD* [42]. Quantitative results are obtained in the PHSPD dataset by computing the 3D

	SURREAL		PHSPD	
	ours (w/o normal)	ours	ours (w/o normal)	ours
polarization image	83.43/94.00	67.25/75.94	84.44/96.42	66.32/74.46
improvement	**16.18/18.06**		**18.12/21.96**	
color image	88.53/100.32	80.70/91.51	85.67/80.34	77.72/70.07
improvement	**7.82/8.81**		**7.95/10.27**	

Table 3. Qualitative ablation study of our SfP approach (w/ vs. w/o the estimated surface normal). MPJPE is the evaluation metric with millimeter unit. Experiments are carried out on both color and polarization images of SURREAL and PHSPD datasets.

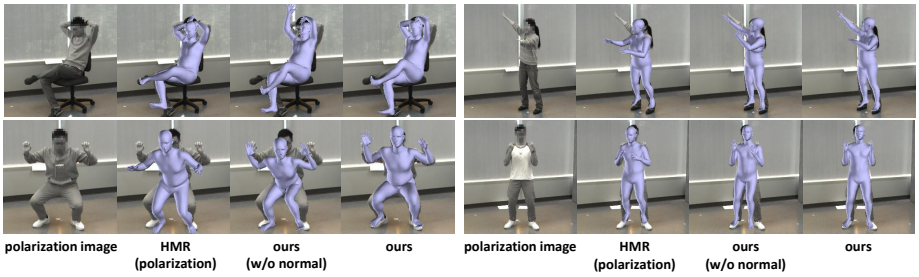

Fig. 4. Exemplar shape estimation results. The first column is polarization images. HMR (polarization) means the HMR model is retrained on polarization images of our PHSPD dataset. Ours (w/o normal) means the model is trained without the normal map as a part of the input.

surface error of the predicted human mesh with respect to the ground-truth mesh. Scaled rigid ICP is performed before the evaluation so as to scale and transform the predicted mesh into the same coordinates as the ground-truth surface. The results are displayed in Table 4. *PIFu* [40] performs the worst, partly as it does not take human pose into consideration when predicting the implicit surface function inside a volume. The 3D surface error from *HMD* [42] and Depth *Human* [41] are relatively small; our SfP approach achieves the best performance, which is partly due to its exploitation of the estimated normal maps. Note the comparison methods of *PIFu* [40], *Depth Human* [41] and *HMD* [42] only work with color images as input. In this experiment, for each of the polarization images used by the two SfP variants, namely *our (w/o deform)* and *ours*, the closet color image captured in the multi-camera setup of PHSPD is taken as its corresponding input to the three comparison methods.

Exemplar visual results are presented in Fig. 5, where the predicted body shapes are overlaid onto the input images. It is observed that the body shapes predicted by *PIFu* and *Depth Human* are generally well-aligned with the input image as they are actually predicting the implicit function value or depth value for each pixel of the foreground human shape. Meanwhile, it does not necessarily indicate accurate alignment of 3D surface mesh, as is evidenced in Table 4. For

	PIFu [40]	Depth Human [41]	HMD [42]	ours (w/o deform)	ours
3D surface error(mm)	73.13	51.02	51.71	41.10	**38.92**

Table 4. Quantitative evaluation of clothed human shape recovery performance methods in the PHSPD dataset.

PIFu and *Depth Human*, the exterior surfaces tend to be overly smooth. Besides, in *Depth Human*, only a partial mesh with respect to the view in the input image is produced. *HMD*, on the other hand, does not work well, as evidenced by the often error-prone surface details. This may be attributed to the less reliable shading representation, given the new environmental lighting and texture ambiguities existed in these color images. Our SfP approach is shown capable of producing reliable prediction of clothed body shapes, which again demonstrates the applicability of polarization imaging in shape estimation, as well as the benefit of engaging the surface normal maps in our approach.

Qualitative results presented in Fig. 6 showcase the robust test performance in novel settings. Note the polarization images are intentionally acquired from unseen human subjects at new geo-locations, so the background scenes are very different from those in the training images.

5 Conclusion

This paper tackles a new problem of estimating clothed human shapes from single 2D polarization images. Our work demonstrate the applicability of engaging polarization cameras as a promising alternative to the existing imaging sensors for human pose and shape estimation. Moreover, by exploiting the rich geometric details in the surface normal of the input polarization images, our SfP approach is capable of reconstructing clothed human body shapes of surface details.

Acknowledgement

This work is supported by the NSERC Discovery Grants, and the University of Alberta-Huawei Joint Innovation Collaboration grants.

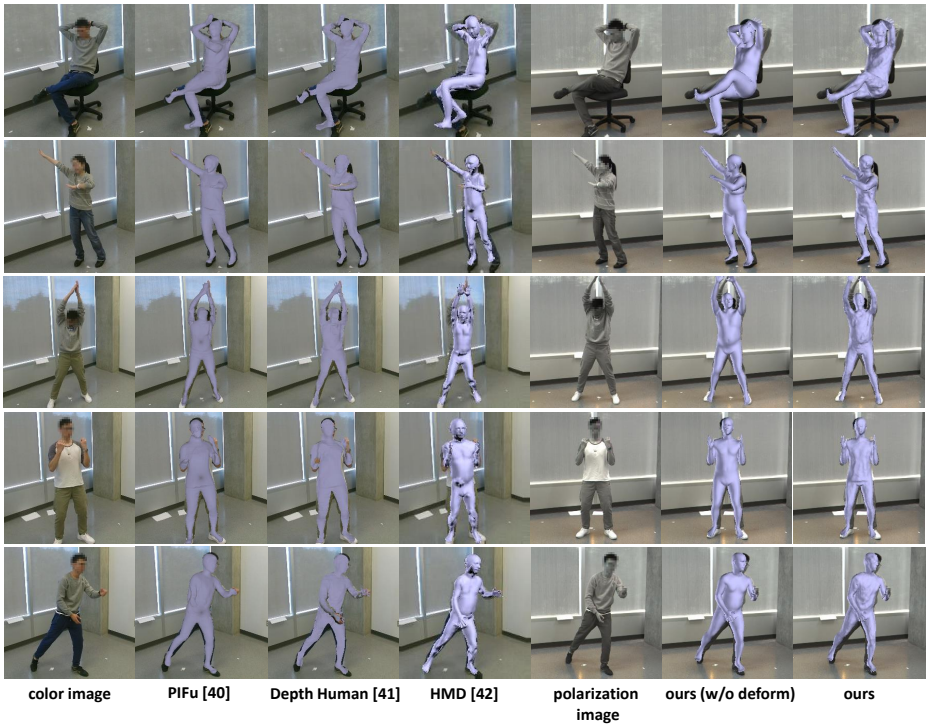

Fig. 5. Exemplar estimation results of clothed body shapes. The first and fifth column are color images and polarization images, respectively. *PIFu* [40], *Depth Human* [41] and *HMD* [42] are the results based on color input images. *Ours (w/o deformation)* and *ours* are the results with the polarization image as the input.

Fig. 6. Exemplar estimation results of clothed body shapes, obtained on polarization images from novel test scenarios (new human subject and scene context).

References

1. Park, S., Hwang, J., Kwak, N.: 3d human pose estimation using convolutional neural networks with 2d pose information. In: European Conference on Computer Vision, Springer (2016) 156–169
2. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2848–2856
3. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3d human pose with deep neural networks. In: British Machine Vision Conference (BMVC). (2016)
4. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: convolutional 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2500–2509
5. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2640–2649
6. Zhao, R., Wang, Y., Martinez, A.M.: A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image. IEEE transactions on pattern analysis and machine intelligence **40**(12) (2017) 3059–3066
7. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2823–2832
8. Nie, B.X., Wei, P., Zhu, S.C.: Monocular 3d human pose estimation by predicting depth on joints. In: Proceedings of the IEEE International Conference on Computer Vision, IEEE (2017) 3467–3475
9. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 398–407
10. Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., Ma, L.: Drpose3d: depth ranking in 3d human pose estimation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. (2018) 978–984
11. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5255–5264
12. Fang, H.S., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018) 6821–6828
13. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7307–7316
14. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 529–545
15. Liu, J., Ding, H., Shahroudy, A., Duan, L.Y., Jiang, X., Wang, G., Chichung, A.K.: Feature boosting network for 3d pose estimation. IEEE transactions on pattern analysis and machine intelligence **42**(2) (2020) 494–501
16. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A., Bangalore, S.B.: Monocular 3d human pose estimation by generation and ordinal ranking. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2325–2334

17. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 10905–10914
18. Wandt, B., Rosenhahn, B.: Repnet: weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7782–7791
19. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9887–9895
20. Wang, K., Lin, L., Jiang, C., Qian, C., Wei, P.: 3d human pose machines with self-supervised learning. IEEE transactions on pattern analysis and machine intelligence (2019)
21. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM transactions on graphics (TOG). Volume 24., ACM (2005) 408–416
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6) (2015) 248
23. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2007) 1–8
24. Dibra, E., Jain, H., Oztireli, C., Ziegler, R., Gross, M.: Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4826–4836
25. Dibra, E., Jain, H., Öztireli, C., Ziegler, R., Gross, M.: Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In: Fourth International Conference on 3D Vision (3DV), IEEE (2016) 108–117
26. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision, Springer (2016) 561–578
27. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6050–6059
28. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7122–7131
29. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 109–117
30. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 459–468
31. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: International Conference on 3D Vision (3DV), IEEE (2018) 484–494
32. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7760–7770

33. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2252–2261
34. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2148–2157
35. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 5349–5358
36. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5614–5623
37. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3d human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3395–3404
38. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: volumetric inference of 3d human body shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 20–36
39. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7739–7749
40. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2304–2314
41. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7750–7759
42. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4491–4500
43. Yang, L., Tan, F., Li, A., Cui, Z., Furukawa, Y., Tan, P.: Polarimetric dense monocular slam. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 3857–3866
44. Ba, Y., Chen, R., Wang, Y., Yan, L., Shi, B., Kadambi, A.: Physics-based neural networks for shape from polarization. arXiv preprint arXiv:1903.10210 (2019)
45. Wehner, R., Müller, M.: The significance of direct sunlight and polarized skylight in the ant's celestial system of navigation. Proceedings of the National Academy of Sciences **103**(33) (2006) 12575–12579
46. Daly, I.M., How, M.J., Partridge, J.C., Temple, S.E., Marshall, N.J., Cronin, T.W., Roberts, N.W.: Dynamic polarization vision in mantis shrimps. Nature communications **7** (2016) 12140
47. Atkinson, G.A., Hancock, E.R.: Recovery of surface orientation from diffuse polarization. IEEE transactions on image processing **15**(6) (2006) 1653–1664
48. Kadambi, A., Taamazyan, V., Shi, B., Raskar, R.: Depth sensing using geometrically constrained polarization normals. International Journal of Computer Vision **125**(1-3) (2017) 34–51
49. Chen, L., Zheng, Y., Subpa-Asa, A., Sato, I.: Polarimetric three-view geometry. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 20–36

50. Cui, Z., Gu, J., Shi, B., Tan, P., Kautz, J.: Polarimetric multi-view stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1558–1567
51. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4966–4975
52. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1446–1455
53. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust estimation of 3d human poses from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2361–2368
54. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In: European Conference on Computer Vision, Springer (2012) 573–586
55. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. IEEE transactions on pattern analysis and machine intelligence **41**(4) (2019) 901–914
56. Zhou, X., Zhu, M., Leonardos, S., Daniilidis, K.: Sparse representation for 3d shape estimation: A convex relaxation approach. IEEE transactions on pattern analysis and machine intelligence **39**(8) (2016) 1648–1661
57. Chen, C.H., Ramanan, D.: 3d human pose estimation = 2d pose estimation + matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7035–7043
58. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2262–2271
59. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2272–2281
60. Jiang, H., Cai, J., Zheng, J.: Skeleton-aware 3d human shape reconstruction from point clouds. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 5431–5441
61. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. ACM transactions on graphics (TOG) **24**(3) (2005) 536–543
62. Zou, S., Zuo, X., Qian, Y., Wang, S., Xu, C., Gong, M., Cheng, L.: Polarization human shape and pose dataset. arXiv preprint arXiv:2004.14899 (2020)
63. Cao, Z., Martinez, G.H., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
64. Zuo, X., Wang, S., Zheng, J., Yu, W., Gong, M., Yang, R., Cheng, L.: Sparsefusion: Dynamic human avatar modeling from sparse rgbd images. IEEE Transactions on Multimedia (2020)
65. Smith, W.A., Ramamoorthi, R., Tozza, S.: Linear depth estimation from an uncalibrated, monocular polarisation image. In: European Conference on Computer Vision, Springer (2016) 109–125

66. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7) (2014) 1325–1339